Reducing Bias in Nonparametric Regression Problems Involving Families of Curves

N. S. Altman *
Biometrics Unit
Cornell University

J. O. Ramsay †
Department of Psychology
McGill University

BU-1175-M

July 1992

Abstract

Nonparametric regression estimators which are linear in the data (linear smoothers) are biased in regions of curvature. Theoretical results such as Stone (1982) place limits on the degree to which the bias can be recovered from a single curve.

However, if the curves were replicated, the bias could be estimated from the regression residuals, and this information used to recenter the estimated curves. Although replicates are seldom available, many problems involve estimation of several curves in the same family. When this is the case, dominant patterns in the residuals should indicate regions in which the estimators are biased. In this paper we show that the average residual can be used to recenter the estimates to improve the total mean squared error of estimation and the coverage of confidence intervals.

Keywords: growth curves; semiparametric regression.

---

*The first author was supported in part by Hatch Grant 151410 NYF and grants from NSERC of Canada and FCAR of Quebec. Much of this work was completed while the first author was visiting the Department of Psychology, McGill University

†The second author was supported in part by grants from NSERC of Canada and FCAR of Quebec.

1 Introduction

In many fields the data collected on the i^{th} sampling unit are determined by curves, $\mu_i(t)$, where t is time or distance. Examples include growth curves, measures of curvature of the spine, and output from monitoring devices such as seismographs. The data are generally contaminated with measurement error. As well, in most cases the data have been collected at discrete design points. The goal of the analysis is generally to recover the underlying curves, or certain features of the curves.

In parametric analysis of such data, the population information is encoded by considering a common functional form for the curves in the family indexed by the unknown parameter. Because such functional forms are often ad hoc, nonparametric regression has been suggested as alternative fitting techniques (Stützle, Gasser, Molinari, Largo, Prader, and Huber, 1980; Gasser, Müller, Köhler, Molinari, and Prader, 1984). However, the nonparametric regression estimators are generally applied to each curve in turn, utilizing little of the population information (except, perhaps, for the selection of tuning parameters).

Linear smoothers, which are among the most commonly used nonparametric regression estimators, are known to be biased in regions in which the underlying functions are curved. The bias is also a function of the design points at which the curves are measured (see, for example, Gasser and Müller 1984 or Jennen-Steinmetz and Gasser 1988). In many families of curves we expect curvature to occur at about the same points on the curve - examples include the timing of the pubertal growth spurt in human height, seasonal effects in weather reports, stock closing in various markets, the location of local extrema in drug response curves. In this paper we show that such common features lead to a population component of bias that can be recovered and used to improve estimation of the individual curves. As well, this recentering improves the coverage of confidence bands around the curves.

The expectation of the regression residual at each design point is the bias of the estimator at that point. If the curves were replicated, the bias of the estimator could be estimated at the design points, using a location estimator. The bias could then be used to reduce the bias of the regression estimate, and recenter confidence intervals.

True replicates are seldom available. However, in many studies a curve is recorded for each sampling unit of a population. Often, only a limited amount of data is available on each sampling unit, but the number of sampling units in the study is large. In this paper, we propose aggregating the residuals from the individual curves to estimate common (population) bias components and show how these estimates may be used to recenter the nonparametric regression estimators to improve the individual curve estimates and recenter confidence intervals. In Section 2, we discuss linear regression with missing variates, and orthogonal series estimators. In Section 3 we use a simulation study to demonstrate the effectiveness of the method both for orthogonal series estimators and other linear smoothers.

2 Recovering Bias Information from Projection Estimators

In this section we show that using the average residual can capture some of the features of missing covariates in finite dimensional regression problems, and provide heuristics for the use of the average residual as an added regressor when using orthogonal series estimators. The argument is made through a logical progression of examples - regression with one missing covariate, regression with a missing finite dimensional vector covariate, and then orthogonal series regression.

In the last case, we think of the underlying true curves as being represented exactly by an infinite expansion in terms of the orthogonal series; however, because the approximation uses only a finite number of terms, we have a missing infinite dimensional covariate which is partially recovered

from the average residual.

Finally, we briefly discuss the use of the average residual to recenter curves based on other linear smoothers which are not orthogonal projections of the data. This last topic is taken up again in the simulation studies in Section 3.

Example 1: Multiple Regression with One Missing Regressor

Example 1 shows that when all the curves are measured at the same fixed design points a single missing regressor which is a function of the design points can be recovered in a certain sense from the residuals averaged pointwise over the curves.

Suppose we have n sampling units, each measured at the same T design points $t' = (t_1, \cdots, t_T)$ and the data, y_{ij}, $i = 1 \cdots n$, $j = 1 \cdots T$ is from the model:

$$y_{ij} = X_j \beta_i + z_j(t)\gamma_i + \varepsilon_{ij}.$$

Here X_j is a p-vector of known variates, $z_j(t)$ is an unknown variate and ε_{ij} are independent and identically distributed with mean 0 and variance σ^2. β_i is a vector of p unknown regression coefficients for each i and γ_i is an unknown constant for each i.

Notice that $z_j(t)$ does not depend on i because it is a function of the fixed, common design points t.

Let X be the matrix with rows X_j, and let z be the column vector with elements $z_j(t)$. For computational convenience, we assume that z is orthogonal to X, that is $z'X = 0$ and $z'z = 1$, where z' denotes the transpose of z. (These restrictions will be removed later.) Also, let the subscript "i." denote the data vector for sampling unit i.

Since only the first p independent variables are known, we compute the estimated least squares regression curve for each sampling unit by:

$$\hat{y}_{i.} = H y_{i.}$$

and the residuals by

$$r_{i.} = (I - H) y_{i.}$$

where $H = X(X'X)^{-1}X'$ and I is the $T \times T$ identity matrix.

Let $\bar{m}$ denote the vector with elements $1/n \sum_{i=1}^{n} m_i$. Then the mean residual

$$\begin{aligned} E(\bar{r}) &= (I - H)E(\bar{y}) \\ &= (I - H)(X\bar{\beta} + z\bar{\gamma} + E(\bar{\varepsilon})) \\ &= z\bar{\gamma} \end{aligned}$$

Note that the expectation here is with respect to the random errors ε. The curves are considered fixed once the sampling units have been selected. The expected average residual is just a multiple of the unknown variate z, and the multiplier depends on the sampling units selected. If z is not orthogonal to X, an orthogonalization argument still gives us that the $Span(X, z) = Span(X, E(\bar{r}))$, where $Span(A, B)$ denotes the vector space spanned by the columns of matrices A and B. So, by regressing y on the augmented matrix $[X|\bar{r}]$ we obtain fitted values $\hat{y}^*$ which are close to unbiased for $E(y)$.

As is well-known, adding a variable reduces bias but increases the variance of the regression estimator. Therefore the mean squared error of the estimator $\hat{y}^*$,

$$MSE(\hat{y}^*) = \sum_{i=1}^{n} \sum_{j=1}^{T} [\hat{y}^*_{i,j} - E(y_{ij})]^2$$

need not be smaller than the mean squared error of $\hat{y}$. As well, in a data analysis situation we must work with $\bar{r}$, rather than $E(\bar{r})$, introducing some additional variation into the estimator. (Of course, this additional variation decreases with sample size.) However, if z is an important predictor, the resulting predicted values are generally an improvement.

Example 2: Multiple Regression with a Missing Finite-Dimensional Vector Regressor

Example 2 shows that when the curves are all measured at the same design points, the principle component of a set of missing regressors, all

functions of the design points, can be recovered in a certain sense from the residuals averaged pointwise over the curves.

Example 1 can be extended to the case where z is replaced by Z, a matrix whose columns are k unknown variates, which depend on t. In this case we have

$$y_{ij} = X_j\beta_i + Z_j(t)\Gamma_i + \varepsilon_{ij},$$

where y_{ij}, X_j and ε_{ij} and β_i are defined as in Example 1. $Z_j(t)$ is the vector of unknown regressors and for each i, Γ_i is a vector of k unknown regression coefficients. Once again, suppose the variables are normalized so that $Z'X = 0$ and $Z'Z = I$.

As in Example 1, we regress $y_{i\cdot}$ on X to obtain fitted values $\hat{y}_{i\cdot}$ and residuals $r_{i\cdot}$ and consider

$$E(\bar{r}) = Z\bar{\Gamma}.$$

$\bar{\Gamma}\bar{\Gamma}'$ is a $k \times k$ symmetric matrix with one non-zero eigenvalue, $\lambda = \bar{\Gamma}'\bar{\Gamma}$. $Z\bar{\Gamma}\bar{\Gamma}'Z'$ therefore also has rank 1, and since

$$Z\bar{\Gamma}\bar{\Gamma}'Z'Z\bar{\Gamma} = \lambda Z\bar{\Gamma}$$

the nonzero eigenvalue is λ with associated eigenvector $E(\bar{r}) = Z\bar{\Gamma}$.

Therefore, $E(\bar{r})$ is a multiple of the largest principal component of the common bias of the sampled units. However, unlike Example 1, when 2 or more regressors are missing, the direction of $E(\bar{r})$ depends on which units were sampled. The component will have the most information when the collection of Γ_i's is most homogeneous.

As in Example 1, the argument does not depend on the orthogonality of X and Z or on the orthonormality of Z. If Z does not satisfy these conditions an orthogonalization argument still gives us that $E(\bar{r})$ is the principal component of common bias of the estimator. So, by regressing y on the augmented matrix $[X|\bar{r}]$ we obtain fitted values $\hat{y}^*$ which are less biased for $E(y)$.

As in Example 1, use of the average residual as an additional regressor reduces bias, but introduces some variance. As well, there is an additional effect since Γ is sample dependent. Improvements in a small, homogeneous subpopulation may be better than those in a larger, more variable sample.

Example 3: Orthogonal series estimators

Example 3 shows that the results of Example 2 can be extended to the case in which a countable number of covariates are missing, each a function of the common design points.

Consider the nonparametric regression model:

$$y_{ij} = \mu_i(t_j) + \varepsilon_{ij},$$

where y_{ij} and ε_{ij} are as in the previous examples, and $\mu_i(t)$ is a function in C_2, the Hilbert space of continuous square integrable functions. Let $\theta_1(t), \theta_2(t), \cdots$ be any orthonormal basis in C_2. Because the data are discrete there is an identifiability problem. There is a unique sequence of real numbers $a_{i,1}, a_{i,2} \cdots$ such that $\mu_i(t) = \sum_{v=1}^{\infty} a_{i,v}\theta_v(t)$. However, for any set of T basis functions, $\theta_{k_1}(t), \cdots, \theta_{k_T}(t)$ there is a sequence of real numbers $b_{k_{i1}}, \cdots, b_{k_{iT}}$ such that $\mu_i(t_j) = \sum_{v=1}^{T} b_{k_{iv}}\theta_{k_v}(t_j)$.

For nonparametric regression problems, the basis is usually ordered in terms of increasing complexity. For example, the two most popular orthogonal series techniques are polynomial regression, for which the basis is ordered by degree, and truncated Fourier analysis, for which the basis is ordered by frequency. The number of basis vectors h to be used in the fit is fixed by some method such as generalized cross-validation (Craven and Wahba 1979). In the discussion of this example, we assume that h is fixed.

Generally, the fit is done by using least squares regression to fit the model

$$y_{ij} = X_j\beta_i + \varepsilon_{ij},$$

where X_j is the row vector $\theta_1(t_j) \cdots \theta_h(t_j)$. Bias is introduced because $h < T$. On the other hand, if T basis vectors are selected, the fit interpolates the data, instead of approximating $\mu_i(t_j)$, so that it is unbiased, although

highly variable, at the design points, and likely quite biased between design points, where we have no means of assessment.

For the purposes of this discussion, we will complete the basis to T elements by adding any $T-h$ functions, $g_{h+1}(t)\cdots g_T(t)$, which are orthogonal to $\theta_1\cdots\theta_h$ and have the orthonormal property $\sum_{j=1}^{T} g_k(t_j)g_v(t_j)=\delta_{k,v}$ where $\delta_{k,v}$ is the Kronecker delta function. Define the matrix Z by the rows $Z_j = g_{h+1}(t_j)\cdots g_T(t_j)$. Then $XZ'=0$ and $ZZ'=I_{T-h}$, and

$$y_{ij} = X_j\beta_i + Z_j\Gamma_i + \varepsilon_{ij}.$$

As in Example 2, we can consider regressing $y_{i\cdot}$ on X to obtain the residuals $r_{i\cdot}$. Then $E(\bar{r})$ is the dominant component of the common bias and this component does not depend on the choice of $g_{h+1}(t)\cdots g_T(t)$.

Example 3 shows that bias reduction can be achieved in orthogonal series estimation by regressing in the average residual just as it can for finite dimensional problems with missing regressors. All that is required is orthogonality of the residuals and fitted values for fixed values of the smoothing parameter h. (The same value of h must be used for every curve in the family.)

Commonly used smoothers such as kernel, nearest neighbor and running linear regression, and smoothing splines are not orthogonal projections. As a result, arguments similar to the heuristics of Example 3 do not apply. However, it seems plausible that similar methodology could be used for bias correction for these estimators. Fitting may be done via the back-fitting algorithm (Hastie and Tibshirani 1990, p. 91). Because the average residual enters the equation linearly, iterative back-fitting is not required, when a linear smoother is used. For fixed smoothing parameter, the smooth estimate can be computed as Sy, where S is the smoother matrix. Denoting the contribution of $\bar{r}$ by $R=\bar{r}(\bar{r}'\bar{r})^{-1}\bar{r}'$, and the $t\times t$ identity matrix by I, then the adjusted estimator is given by Hy where $H=I-(I-S)(I-RS)^{-1}(I-R)$ (Hastie and Tibshirani, 1990, p. 119).

In the next section, the results of a set of simulation studies are discussed.

These show that use of the average residual to correct the curves produces very good results when the initial estimates are badly biased, as they may be if, for example, fits are done with low order polynomials. When smoothing parameters are selected by use of selection criteria such as generalized cross-validation (Craven and Wahba, 1979) the bias and variance of each curve is somewhat balanced. However, when the number of design points is small or the data are highly variable, or when a large number of curves is available, considerable improvements are still possible using either orthogonal series or other smoothing algorithms.

3 Simulations

Three simulation studies were performed to determine how well use of the average residuals improved the fit of the curves and the coverage of the normal theory pointwise confidence intervals.

3.1 Study 1

In the first study, the true curves are generated as fifth degree polynomials. Each simulation result is based on 1000 replicates of samples of 20 curves. The regression coefficients were generated from the multivariate Normal distribution with mean and covariance matrix displayed in Table 1. The parameters were chosen so that the samples of curves would display a number of local maxima and minima in the range $x \in [0, 2]$. 20 equally spaced design points on $[0, 2]$ were used. The fits in this study are all based on fixed numbers of polynomial terms or sine and cosine waves. The 4 simulations in Study 1 are summarized in Table 2.

Goodness of fit was assessed using relative MSE:

$$RelativeMSE = \frac{\sum_i \sum_j (\hat{y}^*_{ij} - E(y_{ij})^2}{\sum_i \sum_j (\hat{y}_{ij} - E(y_{ij})^2} \tag{1}$$

(2)

where $\hat{y}_{ij}$ is the unadjusted (raw) fitted value, and $\hat{y}^*_{ij}$ is the fitted value after adjustment by the average residual.

The relative MSE of the corrected and uncorrected fits are displayed in histograms in Figure 1. Use of the average residual dramatically improves the estimates when the original fitted model is incorrect, and does not introduce much deterioration when the original model was correct. Figure 2 displays the true mean function, fitted cosine curves, and adjusted fits for the some of the curves generated in last replicate of Simulation 3.

Figure 3 displays the average residual for the last replicate of each simulation. In the 3 simulations in which the wrong model was fitted, the average residual shows a marked departure from "random" behavior. In the simulation in which the true model was fitted, the average residual appears to be mainly noise. These plots suggested the use of a "pre-test" estimator, to try to avoid use of the adjustment when it was not needed. However, using an F-test to determine whether or not to include $\bar{r}$ in the estimate did not improve the performance of the estimators.

Figure 4 displays the coverage probabilities for normal theory nominal 95% confidence intervals $\hat{y} \pm t_{20-k}(.05)\sqrt{RMS}h_{ii}$ where $\hat{y}$ is the fitted value, RMS is the residual mean square, h_{ii} is i^{th} the diagonal element of the hat matrix, k is the number of regressor variables, and $t_s(.05)$ is the 5^{th}% percentile of Student's t-distribution on s degrees of freedom. Results for the nominal 90% intervals are similar.

The confidence intervals centered around the unadjusted fits show considerable variability in true coverage probability when the wrong model was fitted - from 100% to less than 10% coverage, with low coverage regions corresponding to regions of high bias in the estimates. The recentered intervals also show some variability, but are much closer to the nominal coverage probabilities, even though the estimate of the error standard deviation is smaller. However, the improvement is not uniform. In simulation using the true model, the unadjusted fit is unbiased for the true mean, and the experimental coverage probabilities are close to their nominal levels. It is

interesting to note that, although the adjustment adds only noise to the estimates, the coverage probabilities of the recentered estimates are only slightly smaller than their nominal levels.

The results of this study show the potential gains using this method. However, it should be noted that, except for the quintic fit, all the fitted models in this study were extremely biased. Lack of fit could readily be detected, even from the residual plot of a single curve. It is therefore not surprising that bias correction performed well.

3.2 Study 2

The second study uses a more realistic set of mean curves and fitting algorithm. To demonstrate the use of orthogonal series estimators, polynomial regression with degree selected by generalized cross-validation were used to fit curves generated from the Bock and Thissen model for human height growth curves (Bock and Thissen, 1980), $\mu_i(t) = 10\sum_{j=1}^{3} A_{ij}/(1 + exp(B_{ij}(t - C_{ij}))$ where t is age in years from 5 to 20 and $\mu_i(t)$ represents height in centimeters. The 9 coefficients for each curve $A_{ij}, B_{ij}, C_{ij}, j = 1 \cdots 3$ were generated to be normally distributed with mean vector and covariance matrix corresponding to those computed by Bock for the Fels Growth Study boys (Bock, personal communication). These are displayed in Table 3. Each simulation result is based on 100 replicates. The 4 simulations in Study 2 are summarized in Table 2.

The median degree selected by GCV was 3 for the high noise cases (simulations 5 and 7) and 7 for the low noise cases. As in other linear smoothing techniques, the fits to the noisier curves are both noisier and more biased. Figure 5 displays a set of 6 mean curves and the unadjusted and adjusted fits for Simulation 5. The polynomial fits are somewhat smoother than the true means, but are qualitatively similar except near the ends. The adjusted fits are less smooth, but pick up the somewhat steeper slope in the teenage years caused by the pubertal growth spurt.

The relative MSE (equation 1) of the adjusted versus polynomial fits

range from 0.8 to 1.1. Gains are greater when the bias, or number of curves are greater. However, even in Simulation 6 (50 curves, $\sigma = 1$) the adjustment provides an improvement in 42/100 trials, and never produces more than 15% deterioration. The average residuals for the high noise trials are roughly sinusoidal with age, while for the low noise trials, the average residuals are small, and appear to be mainly noise.

Figure 6 displays the coverage probabilities of the nominal 95% confidence intervals. For the high noise trials, the coverage of the adjusted intervals is better in the range of ages 10 to 12, near the start of the pubertal growth spurt. For the low noise trials, the coverage of the unadjusted intervals is close to its nominal level, although there is a slight dip in the teenage years. The coverage of the adjusted intervals is slightly worse.

3.3 Study 3

Study 3 was designed to explore the use the average residual in conjunction with a linear smoother which is not a projection operator. Mean curves were generated from same population of quintic mean curves used in Study 1. However, fitting was done using cubic smoothing splines (Wahba, 1975), with smoothing parameter selected subjectively to be .001 for all iterations. The 2 simulations in Study 3 are summarized in Table 2. Each simulation result is based on 100 replicates.

The smoothing spline fits are quite good. Residual plots for individual curves have little information about systematic departures of the fits from the true curves.

Figure 7 displays the true mean, fitted mean, and adjusted fit for 5 of the 100 curves in the final iteration of Simulation 10. The spline smooths provide a very good fit in the center of the interval, but are somewhat biased near the ends. The adjustment adds some noise to the center of the fits, but pick up some of the curvature near the ends. Less obviously, the curvature of the adjusted fits at extrema are closer to the curvature of the true curves. The relative MSE of the adjusted versus unadjusted fits ranges from .73

to 1.13. However, even in the high noise case, in 89 of the 100 trials the adjustment improved the MSE.

Normal theory confidence intervals where computed from the formula $\hat{y} \pm z(\alpha)\hat{\sigma}\sqrt{A_{ii}}$ where $\hat{y}$ is the fitted value or adjusted fitted value, $z(\alpha)$ is the $1-\alpha$ percentile of the standard normal distribution, $\hat{\sigma}^2$ is the Gasser–Sroka–Jennen-Steinmetz variance estimator (Gasser, Sroka, Jennen-Steinmetz, 1986) and A_{ii} is the i^{th} diagonal element of S for the unadjusted estimator and H for the adjusted estimator. Figure 8 shows the simulated coverage probabilities for the nominal 95% confidence intervals. Results for the nominal 90% confidence intervals are similar.

The coverage for the unadjusted intervals is surprisingly poor near the extrema, given that qualitatively the fit appears satisfactory. This is due to the fact that linear smoothers fill in valleys and erode peaks, so that the intervals are not properly centered. The adjusted intervals do much better in the center of the interval. At the ends of the interval, where the adjustment picks up curvature missed by the unadjusted fits, the coverage is actually poorer. Once again, comparing with Figure 8, it appears that the adjusted fits exhibit too much curvature near the ends.

4 Conclusions

In this article we have demonstrated the use of a simple but effective method for improving fit and confidence intervals for nonparametric curves, when a number of similar curves are available. Critical to the success of the method is that features such as local extrema are present at roughly the same location on each curve so that the curves display a common component of bias.

All proofs and simulations have been done using identical design points for each curve. However, the method can readily be extended to use with different design points (on the same interval) by accumulating residuals across curves and smoothing to obtain the "average residual". It is then necessary

to extract the vector of "average residuals" for each set of design points so that the adjustment can be handled.

5 References

Bock, R. D. and Thissen, D. (1980) "Statistical Problems of Fitting Individual Growth Curves." In *Human Physical Growth and Maturation Methodologies and Factors* ed. F. E. Johnston, A. F. Roche, and C. S. Plenum, pp 265-290.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377-403.

Gasser, T., Müller, H-G. (1984) "Estimating Regression Functions and Their Derivatives by the Kernel Method," *Scand. J. Statist.* **11** 171-185.

Gasser, T., Müller, H-G., Köhler, W., Molinari, L. and Prader, A. (1984) Nonparametric regression analysis of growth curves. *Ann. of Stat.* **12** 210-229.

Jennen-Steinmetz, C. and Gasser T. (1988) "A Unifying Approach to Nonparametric Regression Estimation," *Journal of the American Statistical Association*, **83**, 1084-1089.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models.* Chapman and Hall: New York.

Stone, C. J. (1982) "Optimal Global Rates of Convergence for Nonparametric Regression," *Ann. of Stat.* **10** 1040-1053.

Stützle, W., Gasser, Th., Molinari, L., Largo, R.H., Prader, A., and Huber, P.J. (1980) Shape-invariant modeling of human growth. *Ann. Hum. Biol.* **7** 507-528.

Wahba, G. (1975) "Smoothing Noisy Data with Spline Functions", *Nuerische Mathematik* **24**, 383-393.

	a0	a1	a2	a3	a4	a5
mean	-30	20	40	-20	50	20
covariance						
a0	100	25	-30	20	30	-20
a1	25	100	40	80	30	0
a2	-30	40	100	-40	0	50
a3	20	80	-40	100	-60	20
a4	30	30	0	-60	100	30
a5	-20	0	50	20	30	100

Table 1: Mean and covariance of coefficients for quintic polynomials used in Simulation Studies 1 and 3.

Study 1:	True mean - quintic polynomial		
Simulation	number of curves	error standard deviation	method of fit
1	20	1.0	quadratic polynomial
2	20	0.1	quadratic polynomial
3	20	0.1	Fourier polynomial cos x, sin x, cos 2x, sin 2x
4	20	0.1	quintic polynomial
Study 2:	True mean - Bock and Thissen growth curves for human height. Fit is by polynomial regression with degree selected by generalized cross-validation.		
Simulation	number of curves	error standard deviation	method of fit
5	50	0.1	polynomial regression with GCV
6	50	0.4	polynomial regression with GCV
7	100	0.1	polynomial regression with GCV
8	100	0.4	polynomial regression with GCV
Study 3:	True mean function - quintic polynomial. Fitting is done using a cubic smoothing spline with smoothing parameter=.001.		
Simulation	number of curves	error standard deviation	method of fit
9	20	0.01	cubic smoothing spline $\lambda = .001$
10	1000	0.01	cubic smoothing spline $\lambda = .001$

Table 2: Description of Simulations.

	a1	b1	c1	a2	b2	c2	a3	b3	c3
mean	102.76	0.62	-0.53	54.70	0.41	7.80	24.45	1.10	13.73
covariance									
a1	242.24	-5.44	-1.09	-206.36	0.74	19.95	2.87	-0.035	6.54
b1	-5.44	0.14	0.044	4.88	-0.015	-0.49	0.17	-0.0064	-0.15
c1	-1.09	0.044	0.032	1.56	-0.0050	-0.13	0.17	-0.0098	-0.011
a2	-206.36	4.88	1.56	206.80	-0.77	-18.54	-3.29	-0.18	-5.46
b2	0.74	-0.015	-0.0050	-0.77	0.0042	0.054	0.020	0.0012	-0.00042
c2	19.95	-0.49	-0.13	-18.54	0.054	1.88	0.28	0.0022	0.76
a3	2.87	0.17	0.17	-3.29	0.020	0.28	5.70	-0.058	0.037
b3	-0.035	-0.0064	-0.0098	-0.18	0.0012	0.0022	-0.058	0.0067	-0.042
c3	6.54	-0.15	-0.011	-5.46	-0.00042	0.76	0.037	-0.042	0.80

Table 3: Mean and covariance of coefficients for Bock and Thissen models used in Simulation Study 2.

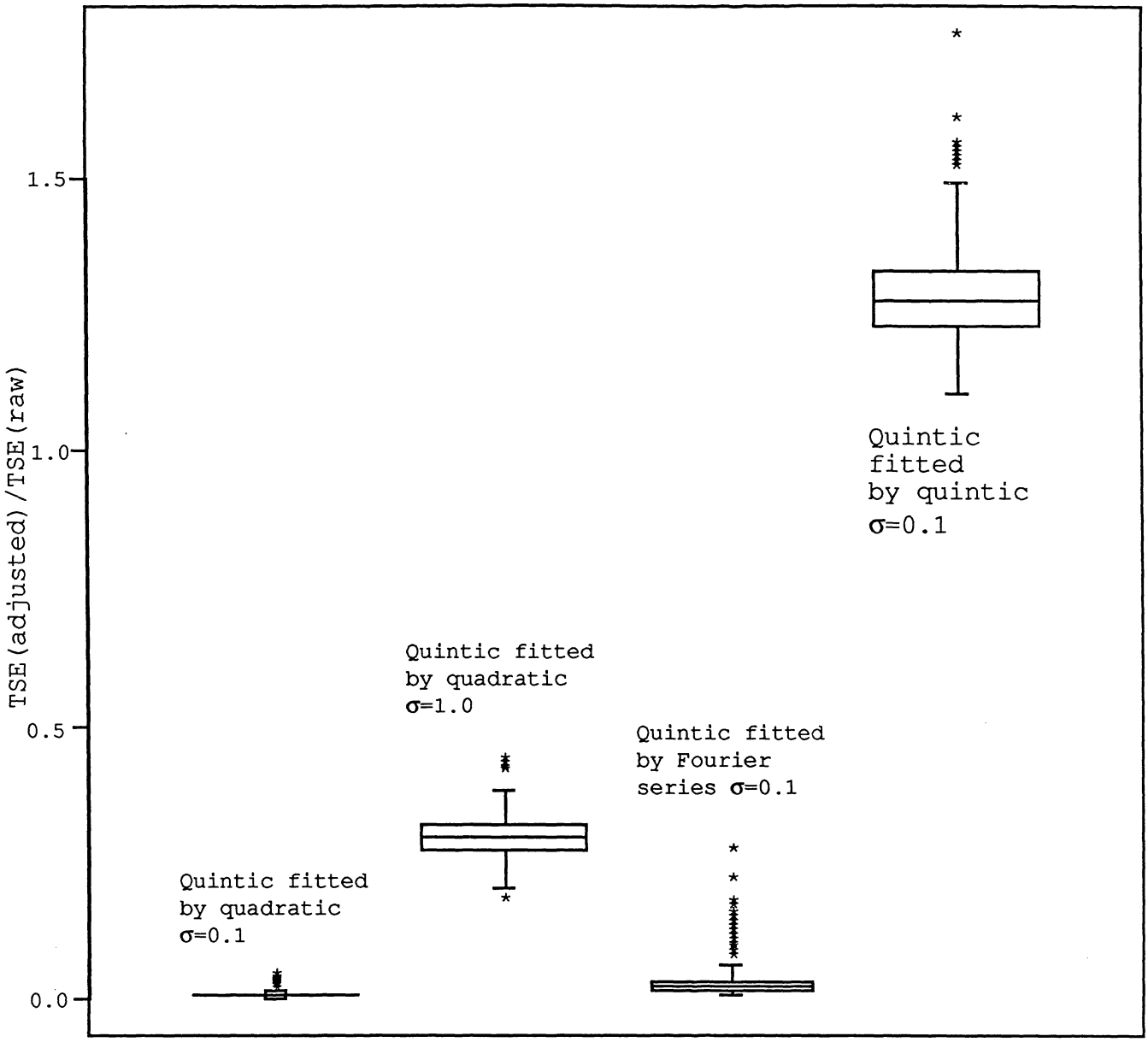


Figure 1: Boxplots of TSE(adjusted)/TSE(raw) fror the 4 simulations of Study 1. Each replicate consists of a set of 20 curves with 20 points per curve. True means were quintic polynomials. Adjusted fits were the raw (polynomial or Fourier) fits with added covariate the mean residual.

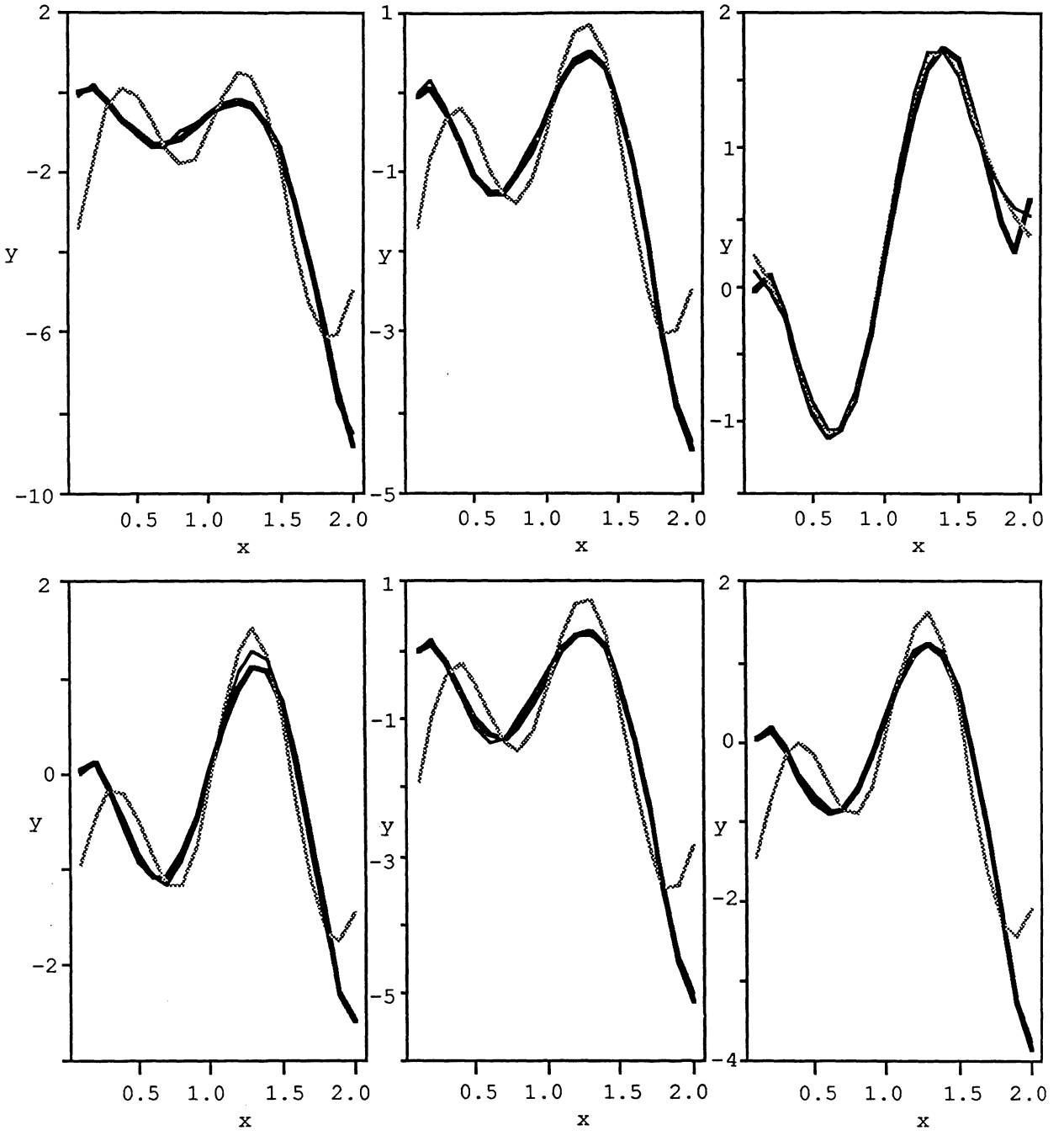


Figure 2: True quintic mean (——), fitted values from 5 term Fourier series (-----) and adjusted fit (——) for 6 of the 20 curves generated in the last replicate of Simulation 3. Notice the excellent bias correction of the adjusted fit.

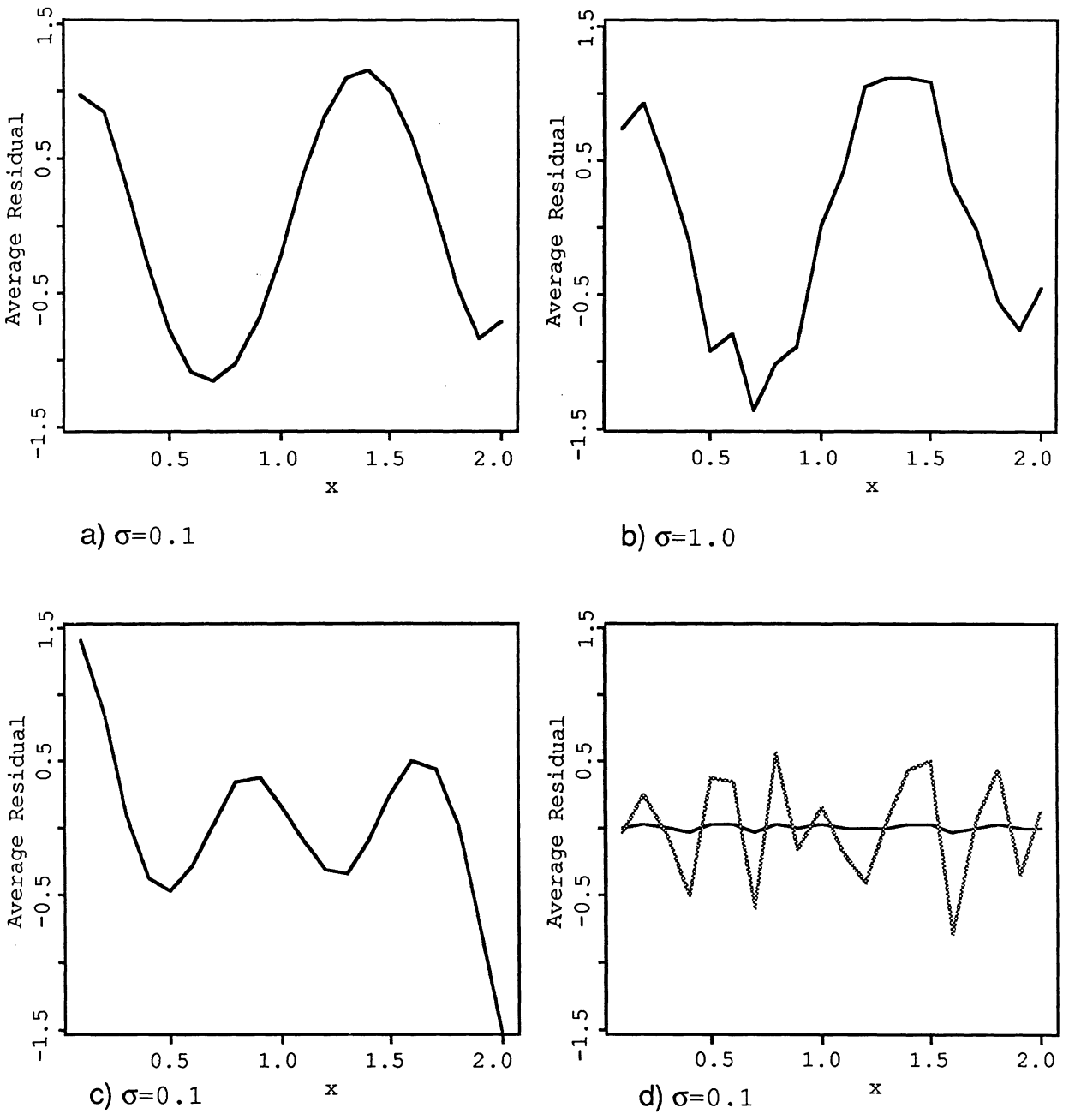


Figure 3: Residual averaged over the 20 curves for the final replicate of Simulations 1-4. True means were quintic polynomials. Raw fits were a) quadratic b) quadratic c) 5 term Fourier series d) quintic. The dashed line in d) is an expanded view of the average residuals, showing the noisy pattern.

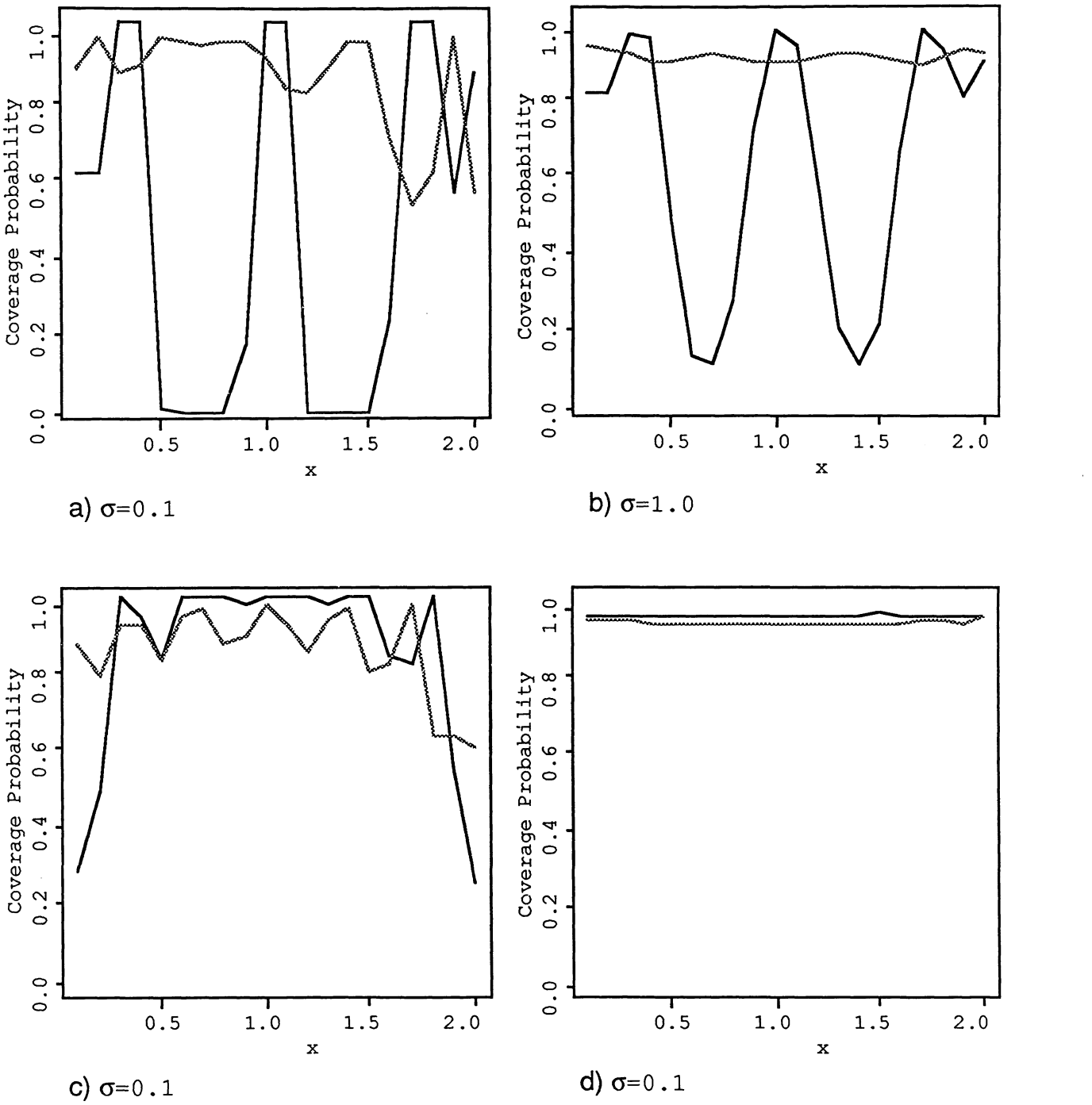


Figure 4: Coverage of Normal theory nominal 95% confidence intervals for each simulation of Study 1. (——) raw fit (----) adjusted fit. True means were quintic polynomials. Raw fits were a) quadratic b) quadratic c) 5 term Fourier series d) quintic. Note that the coverage of adjusted fits for the quintic raw fit is still close to the nominal level. For the other fits, the average coverage is improved, although coverage may be reduced at some design points.

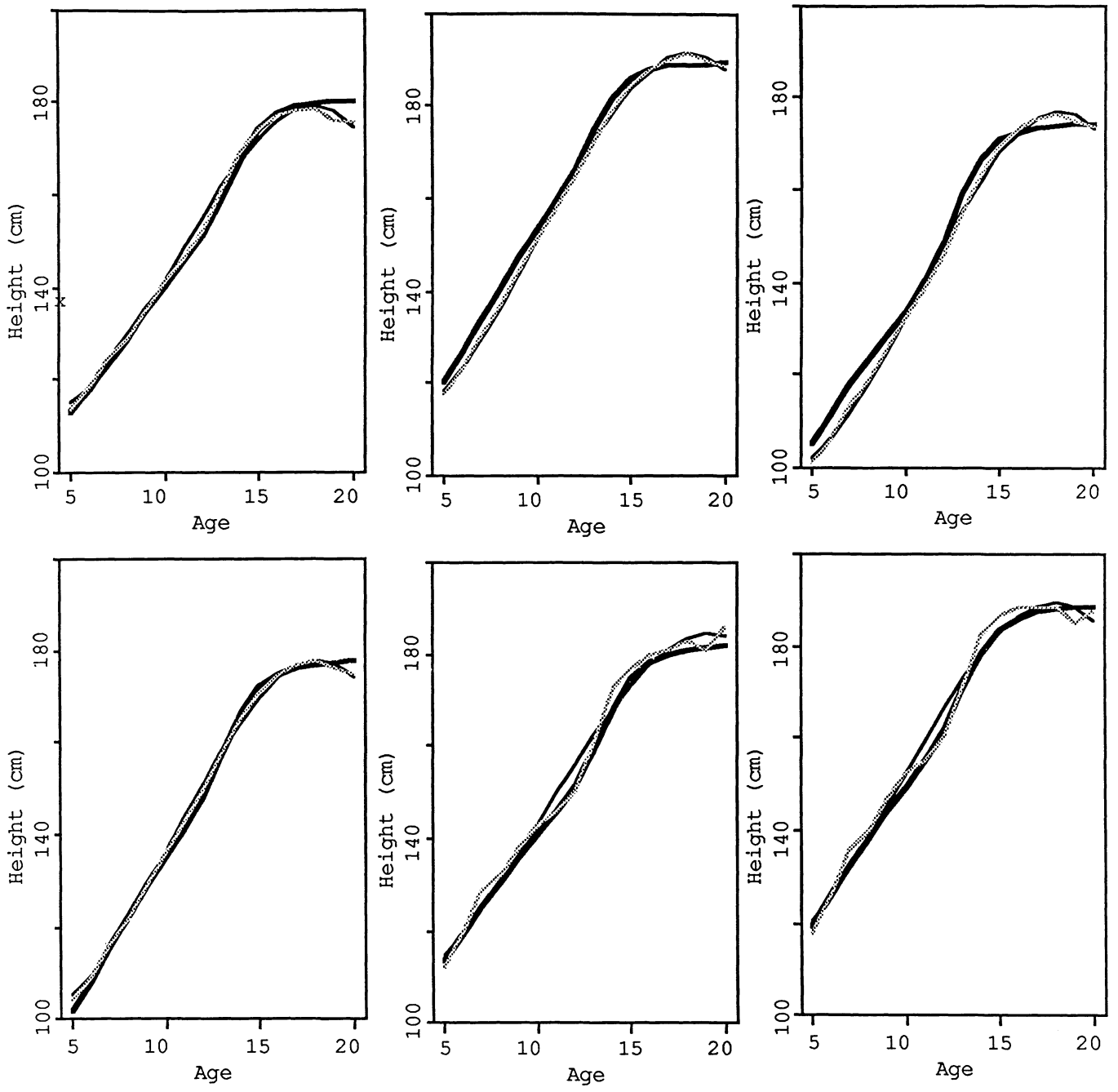
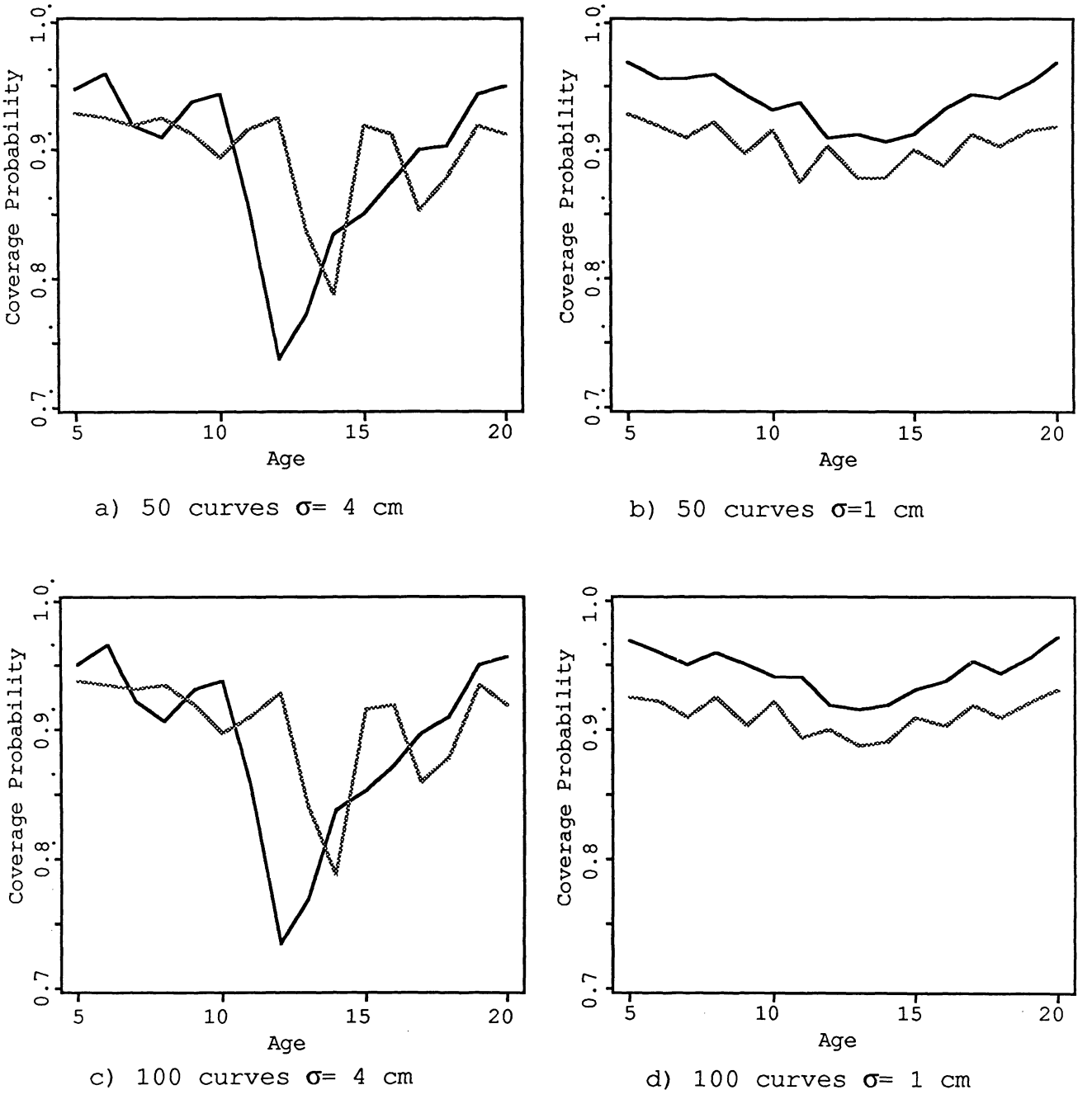


Figure 5: True mean (——), polynomial fit (——) and adjusted fit (----) for simulated growth curves fitted by polynomial regression with degree selected by generalized cross-validation.



a) 50 curves σ= 4 cm

b) 50 curves σ=1 cm

c) 100 curves σ= 4 cm

d) 100 curves σ= 1 cm

Figure 6: Coverage of Normal theory nominal 95% confidence intervals for simulated growth curves fitted by polynomial regression. (——) raw fit (----) adjusted fit. For larger variance, the fitted curves have substantial bias (median selected degree is 3) and so adjustment improves coverage. For smaller variance, the fitted curves have little bias (median selected degree is 7) so adjustment is detrimental to coverage.

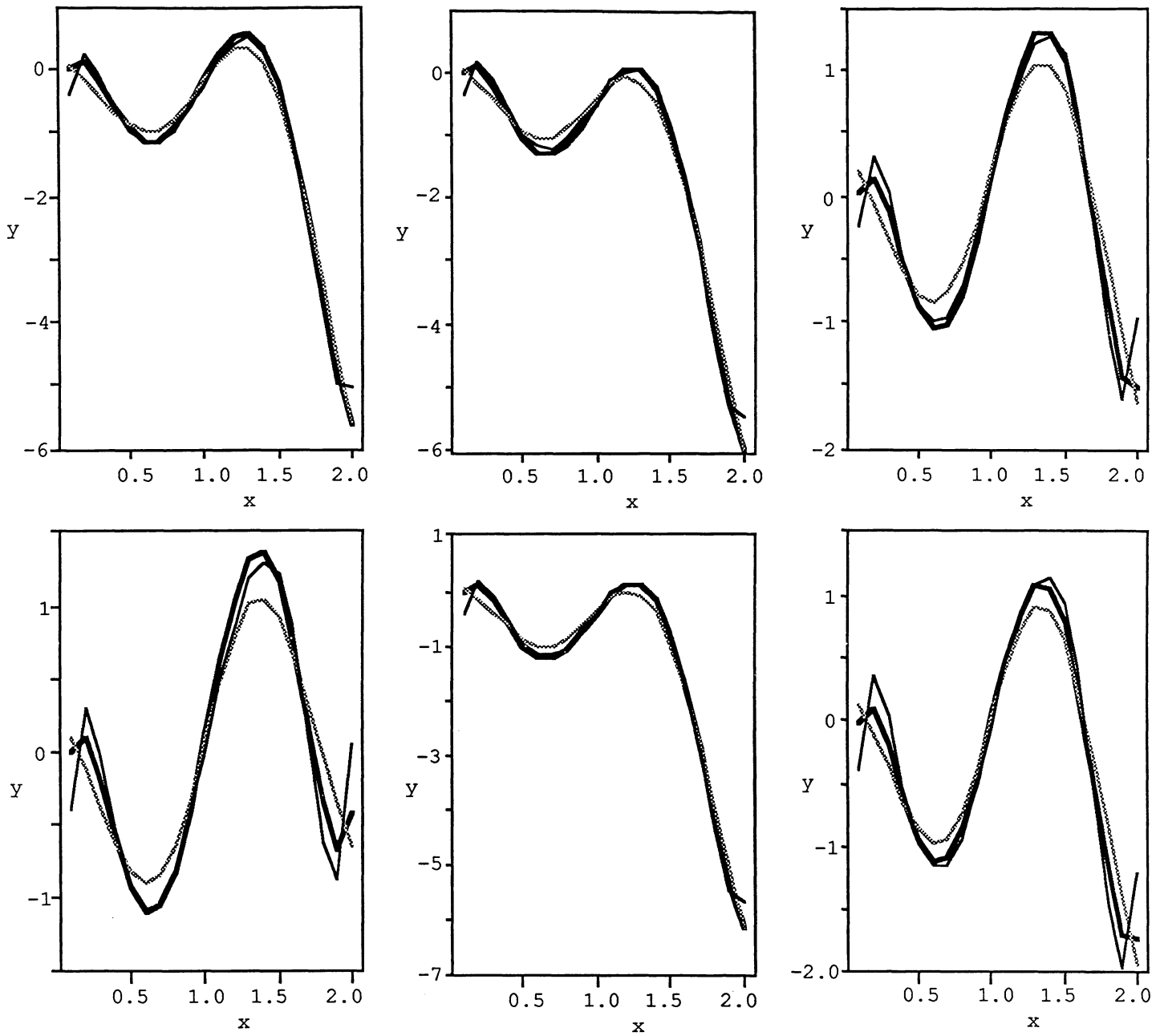


Figure 7: True quintic mean (——), spline fit (-----) and adjusted fit (——) for 6 of the 100 curves generated in the last replicate of Simulation 8. Notice the substantial improvement in fit.

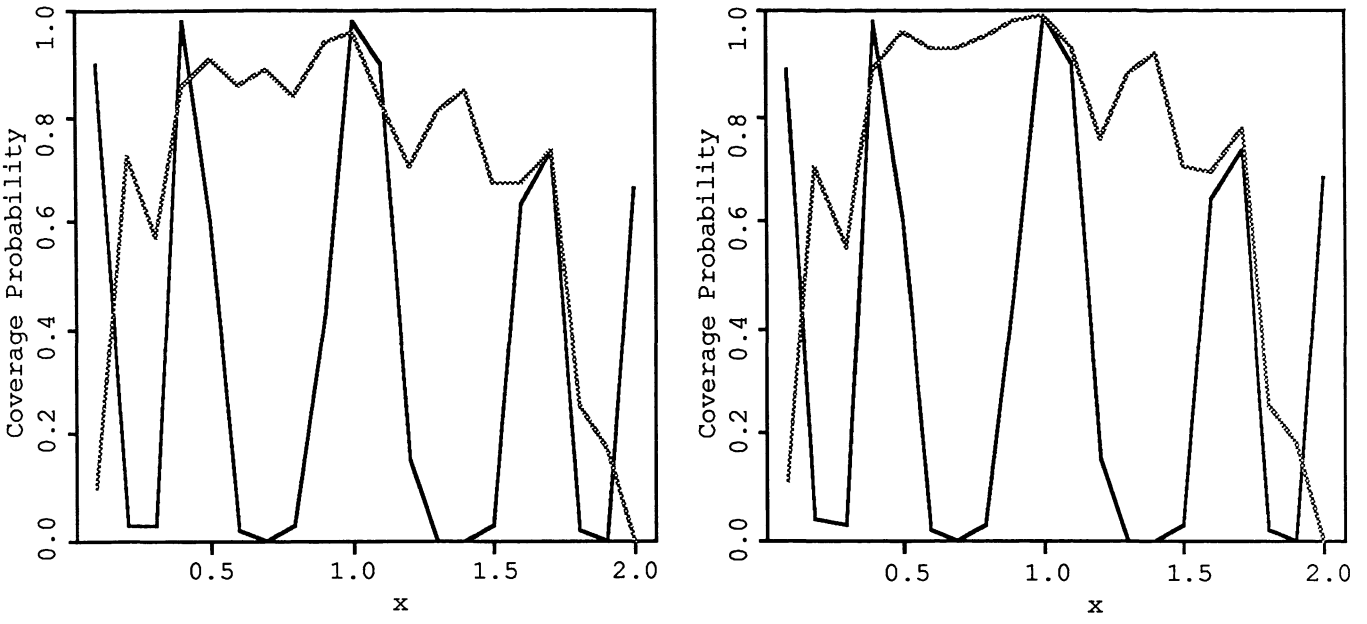


Figure 8: Coverage of Normal theory nominal 95% confidence intervals for quintic polynomials fitted by smoothing splines. raw fit (——) adjusted fit (----). Improvement is substantial near local extrema, but coverage is poor at the ends of the interval.